

Spatial Source Subtraction Based on Incomplete Measurements of Relative Transfer Function

Zbyněk Koldovský^a, Jiří Málek^a, and Sharon Gannot^b

^aFaculty of Mechatronics, Informatics, and Interdisciplinary Studies,
Technical University of Liberec,
Studentská 2, 461 17 Liberec, Czech Republic.
E-mail: {zbynek.koldovsky, jiri.malek}@tul.cz,
fax:+420-485-353112, tel:+420-485-353534

^bFaculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel.
E-mail: Sharon.Gannot@biu.ac.il, fax: 972-3-7384051, tel: 972-3-531-7618

Abstract

Relative impulse responses between microphones are usually long and dense due to the reverberant acoustic environment. Estimating them from short and noisy recordings poses a long-standing challenge of audio signal processing. In this paper we apply a novel strategy based on ideas of Compressed Sensing. Relative transfer function (RTF) corresponding to the relative impulse response can often be estimated accurately from noisy data but only for certain frequencies. This means that often only an incomplete measurement of the RTF is available. A complete RTF estimate can be obtained through finding its sparsest representation in the time-domain: that is, through computing the sparsest among the corresponding relative impulse responses. Based on this approach, we propose to estimate the RTF from noisy data in three steps. First, the RTF is estimated using any conventional method such as the non-stationarity-based estimator by Gannot et al. or through Blind Source Separation. Second, frequencies are determined for which the RTF estimate appears to be accurate. Third, the RTF is reconstructed through solving a weighted ℓ_1 convex program, which we propose to solve via a computationally efficient variant of the SpaRSA (Sparse Reconstruction by Separable Approximation) algorithm. An extensive experimental study with real-world recordings has been conducted. It has been shown that the proposed method is capable of improving many conventional estimators used as the first step in most situations.

Index Terms

Relative Transfer Function, Relative Impulse Response, Sparse Approximations, ℓ_1 norm, Compressed Sensing

I. INTRODUCTION

Noise reduction, speech enhancement and signal separation have been goals in audio signal processing for decades. Although various methods were already proposed and also applied in practice, there still remain open problems. The main reason is that the propagation of sound in a natural acoustic environment is complex. Acoustical signals are wideband in nature and span a frequency range from 20 Hz to 20 kHz. Typical room impulse responses have thousands of coefficients; this aspect makes them difficult to estimate, especially in noisy conditions.

When dealing with, e.g., noise reduction, the crucial question is “what is the unwanted part of the signal to be removed?” Single-channel methods, most of which were developed earlier than multichannel methods, typically rely on some knowledge of noise or interference spectra. For example, the spectra can be acquired during noise-only periods, provided that information about the target source activity is available; see an overview of single-channel methods, e.g., in [1], [2], [3]. Multichannel methods can also use spatial information [3], [4]. For example, a multichannel filter can be designed to cancel the

⁰This work was supported by The Czech Sciences Foundation through Project No. 14-11898S.

signal coming from the target's position. The output of this filter contains only noise and interference components and provides the key reference for the signal enhancement tasks.

Several terms are used in connection with the target signal cancellation, in particular, spatial source subtraction, null beamforming, target cancellation filter, and blocking matrix (BM). The latter refers to one of the building blocks of the minimum variance distortionless (MVDR) beamformer implemented in a generalized sidelobe canceler structure [5]. The BM block is responsible for steering a null towards the desired source, hence blocking, yielding noise-only reference signals, further used to enhance the desired source through adaptive interference canceler and/or by a postfilter. Null beamformers were originally designed under the assumption of free-field propagation (no reverberation) knowing the microphone array geometry (e.g. linear or circular). But later they were also designed taking the reverberation into account; see, e.g., [6], [7].

In natural acoustic environments, the reverberation must be taken into account to achieve satisfactory signal cancellation. This could be done knowing relative impulse responses or, equivalently, relative transfer functions (RTFs) between microphones [6]. The RTF depends on the properties of the environment and on the positions of the target source and microphones. It can be easily computed from noise-free recordings when the target is static [8], [9]. However, the environment as well as the position of the target source can change quickly. Therefore, methods capable of estimating current RTF within short intervals of noisy recordings, during which the target is approximately static, are desirable.

There have been many attempts to estimate the RTF, or (more generally speaking) to design a null beamformer, from noisy recordings [6], [10], [11]. A popular approach is to use Blind Source Separation (BSS) based on Independent Component Analysis (ICA). However, the accuracy of ICA declines with the number of estimated parameters as it is a statistical approach [12]. The blind estimation of the RTF thus poses a challenging problem since there are thousands of coefficients (parameters) to be estimated. The difficulty of this task particularly grows with growing reverberation time and with growing distance of the target source. A recent goal has therefore been to simplify the task through incorporation of prior knowledge. For example, the knowledge of approximate direction-of-arrival of the target is used in [13], [14], or a set of pre-estimated RTFs for potential positions of the target is assumed in [15], [16], [17].

A novel strategy is used in [18], [19], [20] by considering the fact that relative impulse responses can be replaced or approximated by sparse filters, that is, by filters that have many coefficients equal to zero; see also [21], a recent work on sparse approximations of room impulse responses. The authors of [20] propose a semi-blind approach assuming knowledge of the support of a sparse approximation. Hence only nonzero coefficients are estimated using ICA, which implies a significant dimensionality reduction of the parameter space. Results show that sparse estimates of filters achieve better target cancellation than dense filters that are estimated in a fully blind way. However, the assumption that the filter support is known is rather impractical.

In this paper, we propose a novel method based on the idea that the RTF could be known or accurately estimated only in several frequency bins. An appropriate name for such observation is *the incomplete measurement of the RTF*. The entire RTF is then reconstructed by finding a sparse representation of the incomplete measurement in the time-domain. In other words, the relative impulse response between the microphones is replaced by a sparse impulse response whose Fourier transform is, for known frequencies, (approximately) equal to the incomplete RTF. In fact, the idea draws on Compressed Sensing usually applied to sparse/compressible signals or images [22] as well as to system identification.

The following Section introduces the audio mixture model. Section III describes several methods to estimate the relative impulse response or the RTF, both when noise is or is not active. Section IV describes the proposed method, in which the incomplete RTF is reconstructed by an algorithm solving a weighted LASSO program with ℓ_1 sparsity-inducing regularization. Section V then describes several ways to select the incomplete RTF estimate. Section VI presents an extensive experimental study with real recordings, and Section VII concludes this article.

II. PROBLEM DESCRIPTION

A. Model

We will consider situations where two microphones are available¹. A stereo noisy observation of a target signal $s(n)$ can be described as

$$\begin{aligned} x_L(n) &= \{h_L * s\}(n) + y_L(n) \\ x_R(n) &= \{h_R * s\}(n) + y_R(n) \end{aligned} \quad (1)$$

where n is the time index taking values $1, \dots, N$; $*$ denotes the convolution; x_L and x_R are, respectively, the signals from the left and right microphones; and y_L and y_R are the remaining signals (noise and interferences) commonly referred to as noise. Further, h_L and h_R denote the microphone-target acoustical impulse responses. The signals as well as the impulse responses are supposed to be real-valued.

This model assumes that the position of the target source remains (approximately) fixed during the recording interval, i.e., for $n = 1, \dots, N$.

Using the relative impulse response between the microphones denoted as h_{rel} , (1) can be re-written as

$$\begin{aligned} x_L(n) &= s_L(n) + y_L(n) \\ x_R(n) &= \{h_{\text{rel}} * s_L\}(n) + y_R(n) \end{aligned} \quad (2)$$

where $s_L(n) = \{h_L * s\}(n)$ and $h_{\text{rel}} = h_L^{-1} * h_R$ where h_L^{-1} denotes the filter inverse to h_L . Note that although real-world acoustic channels h_L and h_R are causal, h_{rel} need not be so.

The equivalent description of (1) in the short-term frequency-domain is

$$\begin{aligned} X_L(\theta, \ell) &= H_L(\theta)S(\theta, \ell) + Y_L(\theta, \ell), \\ X_R(\theta, \ell) &= H_R(\theta)S(\theta, \ell) + Y_R(\theta, \ell), \end{aligned} \quad (3)$$

where θ denotes the frequency, and ℓ is the frame index. The analogy to (2) is

$$\begin{aligned} X_L(\theta, \ell) &= S_L(\theta, \ell) + Y_L(\theta, \ell) \\ X_R(\theta, \ell) &= H_{\text{RTF}}(\theta)S_L(\theta, \ell) + Y_R(\theta, \ell) \end{aligned} \quad (4)$$

where $S_L(\theta, \ell) = H_L(\theta)S(\theta, \ell)$. Here $H_{\text{RTF}}(\theta)$ denotes the Fourier transform of h_{rel} , which is called the relative transfer function (RTF). It holds that

$$H_{\text{RTF}}(\theta) = \frac{H_R(\theta)}{H_L(\theta)}.$$

With low impact on generality, we assume that H_L does not have any zeros on the unit circle; see the discussion in [6] on page 1619.

B. Spatial Subtraction of a Target Source

When h_{rel} or H_{RTF} are known, an efficient multichannel filter can be designed that cancels the target signal and only pass through noise signals. Consider two-input single-output filter defined as such that its output is

$$z = h * x_L - x_R. \quad (5)$$

According to (2), it holds that

$$z = \underbrace{(h - h_{\text{rel}}) * s_L}_{\text{target signal leakage}} + \underbrace{h * y_L - y_R}_{\text{noise reference}}. \quad (6)$$

¹In this paper, we focus only on the two-microphone scenario due to its comparatively easy accessibility. The idea, however, may be generalized to more microphones.

For $h = h_{\text{rel}}$, the target signal leakage vanishes, and

$$z = h_{\text{rel}} * y_L - y_R. \quad (7)$$

This is the information provided about the noise signals y_L and y_R , which is crucial in signal separation/enhancement or noise reduction applications. For example, the filter defined through (5) serves as the blocking matrix part in systems having the structure of generalized sidelobe canceler, see, e.g., [6], [8], [9], [23], [24], [25].

To complete the enhancement of the noisy signal, many steps still have to be taken, all of which pose other problems. For example, the spectrum of (7) must sometimes be corrected to approach that of the noise in the signal mixture. The noise reduction itself can be done through adaptive interference cancellation (AIC), a task closely related to Acoustic Echo Cancellation (AEC), and/or postfiltering. For the latter, single-channel noise reduction methods could be used once the noise reference is given [26].

However, all the aforementioned enhancement methods suffer from leakage of the target signal into the noise reference (6). This paper is therefore focused on the central problem: finding an appropriate h in (5) so that the blocking effect remains as good as possible.

III. SURVEY OF KNOWN SOLUTIONS

A. Noise-Free Conditions

When a recording of an active target source is available in which no noise is present, the relative impulse response or the RTF can be easily estimated. Such estimates naturally provide good substitutes for h in (5).

1) *Time-domain estimation using least squares:* The mixture model (2) without noise takes on the form

$$\begin{aligned} x_L(n) &= s_L(n), \\ x_R(n) &= \{h_{\text{rel}} * s_L\}(n), \end{aligned}$$

where $n = 1, \dots, N$. Least squares can be used to estimate the first L coefficients of h_{rel} as

$$\mathbf{h}_{\text{LS}} = \arg \min_{\mathbf{h} \in \mathcal{R}^L} \|\mathbf{x}_R - \mathbf{X}_L \mathbf{h}\|_2^2, \quad (8)$$

where \mathbf{h}_{LS} is the vector of L estimated coefficients of h_{rel} , $\mathbf{x}_R = [x_R(1-D), \dots, x_R(N-D)]^T$ where D is an integer delay due to causality, and

$$\mathbf{X}_L = \begin{pmatrix} x_L(1) & 0 & \dots & 0 \\ x_L(2) & x_L(1) & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ x_L(N) & x_L(N-1) & \dots & x_L(N-L+1) \\ 0 & x_L(N) & \dots & x_L(N-L+2) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & x_L(N) \end{pmatrix}.$$

The solution of (8) is

$$\mathbf{h}_{\text{LS}} = (\mathbf{X}_L^T \mathbf{X}_L)^{-1} \mathbf{X}_L^T \mathbf{x}_R = \mathbf{R}^{-1} \mathbf{p}, \quad (9)$$

where

$$\mathbf{R} = \mathbf{X}_L^T \mathbf{X}_L / N, \quad (10)$$

$$\mathbf{p} = \mathbf{X}_L^T \mathbf{x}_R / N. \quad (11)$$

It is worth noting that the Levinson-Durbin algorithm [27] exploiting the Toeplitz structure of \mathbf{R} can be used to compute \mathbf{h}_{LS} for all filter lengths $1, 2, \dots, L$ in $\mathcal{O}(L^2)$ operations. The consistency of the time-domain estimation was studied in [28].

2) *Frequency-domain estimation*: The noise-free recording, in the short-term frequency-domain, takes on the form

$$\begin{aligned} X_L(\theta, \ell) &= S_L(\theta, \ell), \\ X_R(\theta, \ell) &= H_{\text{RTF}}(\theta) S_L(\theta, \ell). \end{aligned}$$

A straightforward estimate of the RTF is given by

$$\hat{H}_{\text{RTF}}(\theta) = \frac{\sum_{\ell} \overline{X_L(\theta, \ell)} X_R(\theta, \ell)}{\sum_{\ell} |X_L(\theta, \ell)|^2}. \quad (12)$$

B. Estimators Admitting Presence of Noise

1) *Frequency-domain estimator using nonstationarity*: A frequency-domain estimator was proposed by Gannot et al. [6]. It admits the presence of noise signals that are stationary or, at least, much less dynamic compared to the target signal; see also [29].

The model (4) can be written as

$$\begin{aligned} X_L(\theta, \ell) &= S_L(\theta, \ell) + Y_L(\theta, \ell) \\ X_R(\theta, \ell) &= H_{\text{RTF}}(\theta) X_L(\theta, \ell) + U(\theta, \ell) \end{aligned} \quad (13)$$

where $U(\theta, \ell) = Y_R(\theta, \ell) - H_{\text{RTF}}(\theta) Y_L(\theta, \ell)$. Note that, in this form, $U(\theta, \ell)$ and $X_L(\theta, \ell)$ are not independent. Let this model be valid for a certain interval during which $H_{\text{RTF}}(\theta)$ is approximately constant, and let the interval be split into P frames. By (13), we have

$$\Phi_{X_R X_L}^p(\theta) = H_{\text{RTF}}(\theta) \Phi_{X_L X_L}^p(\theta) + \Phi_{U X_L}^p(\theta), \quad (14)$$

where $\Phi_{AB}^p(\theta)$ denotes the (cross) power spectral density between A and B during the p th frame.

According to the assumptions of this method (noise is stationary), $\Phi_{U X_L}^p(\theta)$ should be independent of p (thus written without the frame index) and the following set of equations holds

$$\begin{bmatrix} \Phi_{X_R X_L}^1(\theta) \\ \vdots \\ \Phi_{X_R X_L}^P(\theta) \end{bmatrix} = \begin{bmatrix} \Phi_{X_L X_L}^1(\theta) & 1 \\ \vdots & \\ \Phi_{X_L X_L}^P(\theta) & 1 \end{bmatrix} \begin{bmatrix} H_{\text{RTF}}(\theta) \\ \Phi_{U X_L}(\theta) \end{bmatrix}. \quad (15)$$

Now, the estimate of $H_{\text{RTF}}(\theta)$ is obtained by replacing the (cross-)PSDs in (15) by their sample-based estimates and solving the overdetermined system of equations using least squares. Theoretical analyses of bias and variance of this estimator and of the one given by (12) were presented in [29].

2) *Geometric Source Separation (GSS) by [30]*: The method described here was originally designed to blindly separate directional sources whose directions of arrival (DOAs) must be given in advance (known or estimated). The method then makes use of constrained BSS so that the separating filters are kept close to a beamformer that is steering directional nulls in selected directions. We skip details of this method to save space and refer the reader to [30] or to [31] for a shorter description (pages 674–675); see also a modified variant of GSS in [14].

This method can be used for the RTF estimation as follows. Considering two microphones and two sources, one steered direction is selected in the DOA of the target source. The second direction is either the DOA of the (directional) interferer or, in the case of diffused or omnidirectional noise, in a direction that is apart (say 90°) from that of the target source. Let $\mathbf{W}(\theta)$ denote the resulting separating (2×2) transform that is applied to the mixed signals as

$$\mathbf{y}(\theta, \ell) = \mathbf{W}(\theta) \mathbf{x}(\theta, \ell), \quad (16)$$

where $\mathbf{x}(\theta, \ell) = [X_L(\theta, \ell) X_{R,D}(\theta, \ell)]^T$, and $X_{R,D}(\theta, \ell)$ denotes the short-term Fourier transform of $x_R(n - D)$. Ideally, the elements of $\mathbf{y}(\theta, \ell)$ correspond to individual signals in the selected directions.

Let the first row of $\mathbf{W}(\theta)$ be the filter that steers directional null towards the target source, which means that the first element of $\mathbf{y}(\theta, \ell)$ contains only noise signals. The RTF estimate is then given through

$$\hat{H}_{\text{RTF}}(\theta) = -\frac{W_{11}(\theta)}{W_{12}(\theta)}, \quad (17)$$

where $W_{ij}(\theta)$ denotes the ij th element of $\mathbf{W}(\theta)$.

IV. PROPOSED SOLUTION

A. Motivation and Concept

The estimators described above become biased when the assumptions used in their derivations are violated. For example, the bias in (12) depends on the initial Signal-to-Noise Ratio (SNR), which may vary over time and frequency. Assuming that the SNR is sufficiently high for a given frequency, the estimator is good. But when the SNR is low, the estimator's accuracy is also low. Rather than using inaccurate estimates, we can ignore those corresponding to frequencies with low SNR values. We thus arrive at incomplete information about the RTF. That is, the estimate of $H_{\text{RTF}}(\theta)$ is known only for some θ_s .

Based on this idea, our strategy is to construct an appropriate substitute for h in (5) using an incomplete RTF. Typical relative impulse responses are fast decaying sequences, which are compressible in the time-domain, and can thus be replaced by sparse filters [18], [19], [22], [32]. These are derived through finding sparse solutions of a system built up from incomplete information in a different domain: in our case, the frequency-domain [33], [34].

We thus propose a novel method that consists of three parts²:

- 1) Pre-estimation of the RTF from a (noisy) recording.
- 2) Determination of a subset of frequencies where the estimate of the RTF is sufficiently accurate.
- 3) Computation of a sparse approximation of h_{rel} using the incomplete RTF.

Various solutions can be used for each part. Potential methods to solve Part 1 have been already described in Section III. Part 2 can be solved in many ways depending on a given scenario, signal characteristics and the method used within Part 1; we postpone this issue to the next Section. Now we focus on a mathematical description of an appropriate method to solve Part 3.

B. Nomenclature and Problem Formulation for Part 3

Consider the Discrete Fourier Transform (DFT) domain where the length of the DFT is M (sufficiently large with respect to the effective length of h_{rel}), and, for simplicity, let M be even. Let \mathcal{S} denote the set of indices of frequency bins where a given RTF estimate, denoted as $\hat{H}_{\text{RTF}}(\theta_k)$, $k \in \mathcal{S}$ is sufficiently accurate (that is, assume that Part 1 and 2 have already been resolved). Specifically, let the values of the estimate be

$$\hat{H}_{\text{RTF}}(\theta_k) = f_k, \quad k \in \mathcal{S} \subseteq \{1, \dots, M/2\}, \quad (18)$$

where $\theta_k = 2k\pi/M$.

For simplicity, the frequency bins $k = 0$ and $k = M/2 + 1$ can be excluded from \mathcal{S} for the following symmetry to hold: Once $k \in \mathcal{S}$, then the RTF estimate is also known for θ_{M-k} , namely $\hat{H}_{\text{RTF}}(\theta_{M-k}) = \overline{f_k}$ (the conjugate value of f_k), since h_{rel} is real-valued.

Let \mathbf{h}_{rel} denote an $M \times 1$ column vector stacking M coefficients of h_{rel} , and $\mathbf{f} = [f_1, \dots, f_{|\mathcal{S}|}]^T$ where $|\mathcal{S}|$ denotes the cardinality of \mathcal{S} . The known estimates of the RTF satisfy

$$\mathbf{f} = \mathbf{F}_{\mathcal{S}} \cdot \mathbf{h}_{\text{rel}} \quad (19)$$

²The proposed method can be modified in many ways since various solutions can be used for each part of it. We could therefore speak about a proposed class of methods. Nevertheless, the term ‘‘proposed method’’ will be used throughout the article.

where \mathbf{F} is the $M \times M$ matrix of the DFT, and $\mathbf{F}_{\mathcal{S}}$ is a submatrix of \mathbf{F} comprised of rows whose indices are in \mathcal{S} . Since \mathbf{h}_{rel} is real, the system of linear equations (19) can be written as $2|\mathcal{S}|$ real-valued linear conditions

$$\mathbf{f} = \mathbb{F}_{\mathcal{S}} \cdot \mathbf{h}_{\text{rel}} \quad (20)$$

where $\mathbf{f} = [\Re(\mathbf{f})^T \Im(\mathbf{f})^T]^T$ and $\mathbb{F}_{\mathcal{S}} = [\Re(\mathbf{F}_{\mathcal{S}})^T \Im(\mathbf{F}_{\mathcal{S}})^T]^T$, and $\Re(\cdot)$ and $\Im(\cdot)$ denote, respectively, the real and imaginary parts of the argument.

Since $|\mathcal{S}|$ is typically smaller than $M/2$, the system (20) is underdetermined and has many solutions. The key idea is to find sparse solutions that yield efficient sparse approximations of \mathbf{h}_{rel} .

C. Sparse solutions of (20)

The sparsest solution of (20) is defined as

$$\mathbf{g}_0 = \arg \min_{\mathbf{h}} \|\mathbf{h}\|_0 \quad \text{w.r.t.} \quad \mathbf{f} = \mathbb{F}_{\mathcal{S}} \mathbf{h}, \quad (21)$$

where $\|\mathbf{h}\|_0$ is equal to the number of nonzero elements in \mathbf{h} (the ℓ_0 pseudonorm). Solving this task is an NP-hard problem. Further in the paper, we will therefore consider relaxed variants based on convex programming. Several efficient greedy algorithms to solve (21) exist but cannot guarantee the finding of a global solution in general; see, e.g., [35], [36].

A more tractable formulation is based on the replacement of the ℓ_0 pseudonorm in (21) by ℓ_1 -norm, a sparsity-inducing criterion with that the optimization program becomes convex. The program is called basis pursuit [37] and is defined as

$$\mathbf{g}_{\text{BP}} = \arg \min_{\mathbf{h}} \|\mathbf{h}\|_1 \quad \text{w.r.t.} \quad \mathbf{f} = \mathbb{F}_{\mathcal{S}} \mathbf{h}. \quad (22)$$

Using the substitution $\mathbf{h} = \mathbf{h}^+ - \mathbf{h}^-$ where $\mathbf{h}^+ \geq 0$ and $\mathbf{h}^- \geq 0$, (22) can be recasted as

$$\{\mathbf{g}_{\text{BP}}^+, \mathbf{g}_{\text{BP}}^-\} = \arg \min_{\mathbf{h}^+, \mathbf{h}^-} \mathbf{1}^T (\mathbf{h}^+ + \mathbf{h}^-) \quad (23)$$

under the constraints

$$\mathbf{f} = \mathbb{F}_{\mathcal{S}} (\mathbf{h}^+ - \mathbf{h}^-), \quad \mathbf{h}^+ \geq 0, \quad \mathbf{h}^- \geq 0,$$

which is indeed a linear programming problem. The solution can be found using the standard Matlab `linprog` function. Other state-of-the-art optimization tools can also be used, such as the SPGL1 package³ by Berg et al.; see [38].

However, neither formulation (21) nor (22) takes into account the fact that \mathbf{f} contains certain estimation errors. It is therefore better to relax the constraint given through (20). One such alternative to (22) is LASSO (Least Absolute Shrinkage and Selector Operator) defined as

$$\mathbf{g}_{\text{LASSO}} = \arg \min_{\mathbf{h}} \|\mathbb{F}_{\mathcal{S}} \mathbf{h} - \mathbf{f}\|_2^2 + \tau \|\mathbf{h}\|_1, \quad (24)$$

where $\tau \geq 0$. This formulation is closely related to the basis pursuit denoising program defined as

$$\mathbf{g}_{\text{BPDN}} = \arg \min_{\mathbf{h}} \|\mathbf{h}\|_1 \quad \text{w.r.t.} \quad \|\mathbb{F}_{\mathcal{S}} \mathbf{h} - \mathbf{f}\|_2^2 \leq \epsilon \quad (25)$$

with $\epsilon \geq 0$, which is easy to interpret: The constraint $\|\mathbb{F}_{\mathcal{S}} \mathbf{h} - \mathbf{f}\|_2^2 \leq \epsilon$ is a relaxation of $\mathbf{f} = \mathbb{F}_{\mathcal{S}} \mathbf{h}$ taking the possible inaccuracy in \mathbf{f} into account. LASSO is equivalent to (25) in the sense that the sets of solutions for all possible choices of τ and ϵ are the same. It means that the solution of (25) can be found through solving (24) with the corresponding τ . Nevertheless, the correspondence between τ and ϵ is not trivial and is possibly discontinuous [39].

³<http://www.cs.ubc.ca/~mpf/spgl1>

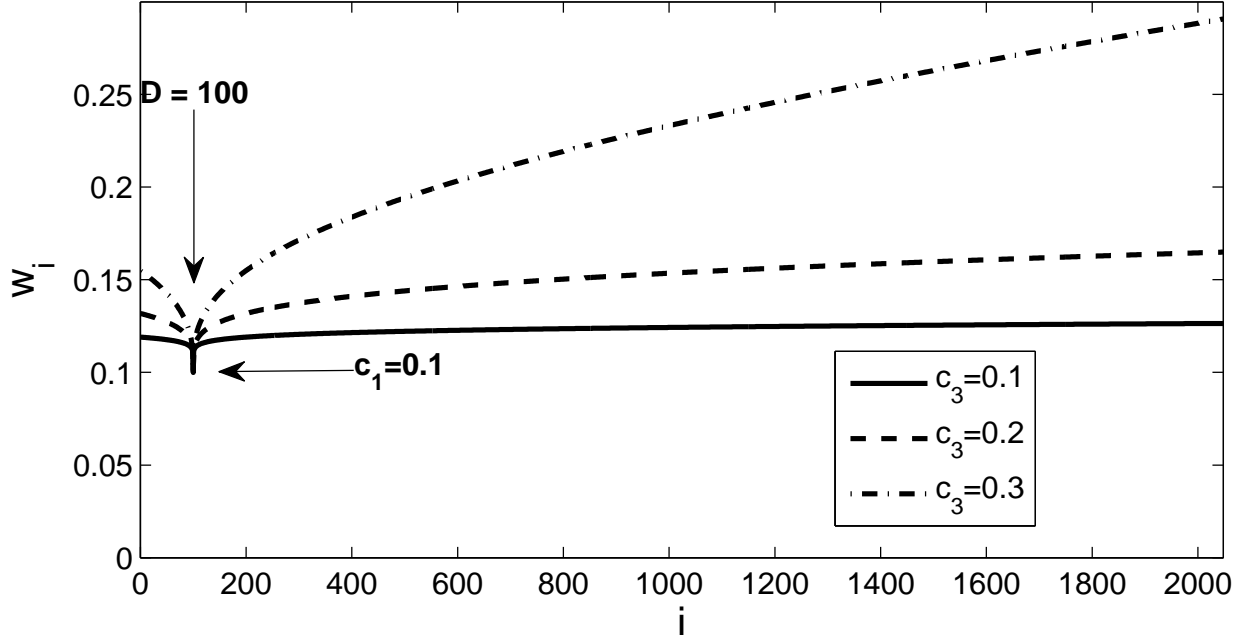


Fig. 1. Example of the weighting function (27) with $M = 2048$, $D = 100$, $c_1 = 0.1$ and $c_2 = 0.11$.

In this paper, we use a weighted formulation of (24) given by

$$\mathbf{g}_{\text{WLASSO}} = \arg \min_{\mathbf{h}} \|\mathbb{F}_{\mathcal{S}} \mathbf{h} - \mathbf{f}\|_2^2 + \|\mathbf{w} \odot \mathbf{h}\|_1, \quad (26)$$

where $\mathbf{w} = [w_1, \dots, w_M]^T$ is a vector of nonnegative weights (absorbing τ), and \odot denotes the element-wise product.

The weights enable us to incorporate a priori knowledge about the solution. Elements of $\mathbf{g}_{\text{WLASSO}}$ with higher weights tend to be closer to or equal to zero. We use this fact and select the weights to reflect the expected shape of h_{rel} . Our heuristic choice, which is similar to that in [21], is

$$w_i = c_1 \cdot e^{c_2|i-D|^{c_3}}, \quad i = 1, \dots, M, \quad (27)$$

where c_j , $j \in \{1, 2, 3\}$, are positive constants. Fig. 1 shows three examples of this weighting function with three different values of the exponent parameter c_3 when $M = 2048$, $D = 100$, $c_1 = 0.1$ and $c_2 = 0.11$. The smallest weights are concentrated near $i = D$, because the direct-path peak of h_{rel} is expected there; the minimum value is $w_D = c_1$. The weights grow with the distance from $i = D$, where the speed of the growth is controlled through c_2 and c_3 . The growth of weights should reflect the expected decay in magnitudes of coefficients in h_{rel} .

D. Algorithm

In this subsection, a proximal gradient algorithm to solve (26) is proposed. It is a modification of SpaRSA (Sparse Reconstruction by Separable Approximation) introduced in [40]; see also closely related iterative shrinkage/thresholding methods [41]. An advantage of these methods is their fast convergence, especially when they are initialized in the vicinity of the solution. The computational load is reduced using the properties of $\mathbb{F}_{\mathcal{S}}$.

Proximal gradient methods could be seen as a generalization of gradient descent algorithms for convex minimization programs where the objective function has the form

$$f(\mathbf{h}) + g(\mathbf{h}),$$

where both f and g are closed proper convex and f is differentiable [42]. Indeed, (26) obeys this form where $f(\mathbf{h}) = \|\mathbb{F}_{\mathcal{S}}\mathbf{h} - \mathbf{f}\|_2^2$ and $g(\mathbf{h}) = \|\mathbf{w} \odot \mathbf{h}\|_1$.

One iteration of the proximal gradient method is

$$\mathbf{h} \leftarrow \text{prox}_{\lambda g}(\mathbf{h} - \lambda \nabla f(\mathbf{h})) \quad (28)$$

where

$$\text{prox}_{\lambda g}(\mathbf{h}) = \arg \min_{\mathbf{z}} (g(\mathbf{z}) + 1/(2\lambda)\|\mathbf{z} - \mathbf{h}\|_2^2) \quad (29)$$

is the proximal operator, and $\lambda > 0$ is a step-length parameter. The method is known to converge under very mild conditions; see [42].

By putting f and g from (26) into (28), we arrive at one iteration of the proposed algorithm

$$\mathbf{h}^{t+1} = \arg \min_{\mathbf{z}} \frac{1}{2}\|\mathbf{z} - \mathbf{u}^t\|_2^2 + \alpha^t \|\mathbf{w} \odot \mathbf{z}\|_1 \quad (30)$$

where $t = 0, 1, 2, \dots$ is the iteration index, $\alpha^t \in [\alpha_{\min}, \alpha_{\max}]$ is a variable step-length parameter, and

$$\mathbf{u}^t = \mathbf{h}^t - \alpha^t \mathbb{F}_{\mathcal{S}}^T (\mathbb{F}_{\mathcal{S}} \mathbf{h}^t - \mathbf{f}). \quad (31)$$

The elements of \mathbf{z} are separable in (30), which allows us to find the solution in closed-form [40], that is

$$\mathbf{h}^{t+1} = \text{soft}(\mathbf{u}^t, \alpha^t \mathbf{w}) \quad (32)$$

where $\text{soft}(u, a) = \text{sign}(u) \max\{|u| - a, 0\}$. In (32), this “soft-thresholding” function is applied element-wise.

The step-length parameter α^t is chosen as in SpaRSA

$$\alpha^t = \frac{\|\mathbf{h}^t - \mathbf{h}^{t-1}\|_2^2}{\|\mathbb{F}_{\mathcal{S}}(\mathbf{h}^t - \mathbf{h}^{t-1})\|_2^2}, \quad (33)$$

which was derived based on a variant of the Barzilai-Borwein spectral approach; see [40].

To terminate the algorithm, we derive a stopping criterion as follows. It holds that $\mathbf{g}_{\text{WLASSO}}$ is the solution of (26) if and only if it satisfies [43]

$$(\mathbb{F}_{\mathcal{S}})_{\Gamma}^T (\mathbb{F}_{\mathcal{S}} \cdot \mathbf{g}_{\text{WLASSO}} - \mathbf{f}) = -\mathbf{w}_{\Gamma} \odot \mathbf{q}_{\Gamma}, \quad (34)$$

$$|(\mathbb{F}_{\mathcal{S}})_{\Gamma^c}^T (\mathbb{F}_{\mathcal{S}} \cdot \mathbf{g}_{\text{WLASSO}} - \mathbf{f})| < \mathbf{w}_{\Gamma^c}, \quad (35)$$

where the subscript $(\cdot)_{\Gamma}$ denotes the restriction to indices (columns in the case of a matrix) in the set Γ ; \mathbf{q} is the vector of signs of $\mathbf{g}_{\text{WLASSO}}$, that is $\mathbf{q} = \text{sign}(\mathbf{g}_{\text{WLASSO}})$; Γ is the set of indices of nonzero elements of $\mathbf{g}_{\text{WLASSO}}$ (the active set), and Γ^c is its complement to $\{1, \dots, M\}$. We define the termination criterion that assesses the degree of validity of (34) as

$$\text{crit}(\mathbf{h}^t) = \left\| (\mathbb{F}_{\mathcal{S}}^T (\mathbb{F}_{\mathcal{S}} \cdot \mathbf{h}^t - \mathbf{f}) + \mathbf{w} \odot \text{sign}(\mathbf{h}^t))_{\Gamma} \right\|_2^2. \quad (36)$$

The algorithm stops iterating when $\text{crit}(\mathbf{h}^t) \leq \text{tol}$ where tol is a small positive constant.

Using the fact that $\mathbf{g}_{\text{WLASSO}}$ satisfies (34) and (35), it can be shown that $\mathbf{g}_{\text{WLASSO}}$ is a fixed point of (30). The global convergence of the algorithm (although with a different stopping criterion) was proven in [40].

Most of the computational burden is due to the vector-matrix products by $\mathbb{F}_{\mathcal{S}}$ and $\mathbb{F}_{\mathcal{S}}^T$ in (31) and in (33). Since $\mathbb{F}_{\mathcal{S}}$ only represents a part of the DFT, the products can be computed via the (inverse) Fast Fourier transform, which also leads to memory savings as $\mathbb{F}_{\mathcal{S}}$ is determined only through \mathcal{S} . The computational complexity of one iteration is thus $\mathcal{O}(M \log M)$. A pseudo-code of the algorithm⁴ is summarized in Algorithm 1.

⁴The Matlab implementation of Algorithm 1 is available at <http://itakura.ite.tul.cz/zbynek/downloads.htm>

Algorithm 1: Algorithm to solve (26)

Input: $\mathcal{S}, \mathbf{f}, \mathbf{w} = [w_1, \dots, w_M]^T, \mathbf{h}^0$
Output: \mathbf{h}^t
 $\mathbf{d} = \mathbf{0}_{M \times 1}, \mathbf{r}^0 = \mathbb{F}_{\mathcal{S}} \mathbf{h}^0 - \mathbf{f}, \nabla^0 = \mathbb{F}_{\mathcal{S}}^T \mathbf{r}^0, i = \sqrt{-1}$
 $t = 0$
while $\text{crit}(\mathbf{h}^t) > \text{tol}$ **do**
 $\mathbf{h}^{t+1} = \text{soft}(\mathbf{h}^t - \alpha \nabla^t, \alpha \mathbf{w})$
 $\Delta \mathbf{h} = \mathbf{h}^{t+1} - \mathbf{h}^t$
 $\mathbf{a} = \text{fft}(\Delta \mathbf{h})$
 $\mathbf{b} = [\Re(\mathbf{a}_{\mathcal{S}})^T \Im(\mathbf{a}_{\mathcal{S}})^T]^T$
 $\mathbf{r}^{t+1} = \mathbf{r}^t + \mathbf{b}$
 $\alpha^{t+1} = \min(\alpha_{\max}, \max(\alpha_{\min}, \|\Delta \mathbf{h}\|_2^2 / \|\mathbf{b}\|_2^2))$
 $\mathbf{d}_{\mathcal{S}} = \mathbf{r}_{1:|\mathcal{S}|}^{t+1} + i \mathbf{r}_{|\mathcal{S}|+1:2|\mathcal{S}|}^{t+1}$
 $\nabla^{t+1} = M/2 \cdot \text{ifft}(\mathbf{d}, M, \text{'symmetric'})$
 $t \leftarrow t + 1$
end

/* now $\mathbf{b} = \mathbb{F}_{\mathcal{S}} \Delta \mathbf{h}$ */
 /* now $\mathbf{r}^{t+1} = \mathbb{F}_{\mathcal{S}} \mathbf{h}^{t+1} - \mathbf{f}$ */

 /* now $\nabla^{t+1} = \mathbb{F}_{\mathcal{S}}^T \mathbf{r}^{t+1}$ */

V. DETERMINING THE SET \mathcal{S}

This Section is dedicated to solutions of Part 2 of the proposed method. Let the estimates $\hat{H}_{\text{RTF}}(\theta_k)$ of $H_{\text{RTF}}(\theta_k)$ be given for all k . The task is to select the set \mathcal{S} such that $\hat{H}_{\text{RTF}}(\theta_k)$ is sufficiently accurate for $k \in \mathcal{S}$.

A. Oracle Inference

For experimental purposes, we define an oracle method that comes from complete knowledge of the SNR in the frequency domain. For simplicity, we can consider the SNR on the left microphone only, which is given by

$$\text{SNR}_{\text{L}}(\theta_k) = \frac{\sum_{\ell} |S_{\text{L}}(\theta_k, \ell)|^2}{\sum_{\ell} |Y_{\text{L}}(\theta_k, \ell)|^2}.$$

This method selects frequencies for which the SNR is higher than a positive adjustable parameter β . The resulting set will be denoted as $\mathcal{S}_{\beta}^{\text{or}}$. Specifically, it holds that

$$k \in \mathcal{S}_{\beta}^{\text{or}} \iff \text{SNR}_{\text{L}}(\theta_k) > \beta. \quad (37)$$

Now we focus on methods that do not require prior knowledge of SNR.

B. Kurtosis-Based Selection

For cases where the target signal is a speaker's voice while the other sources are non-speech, voice activity detectors (VAD) can be used to infer high-SNR frequency bins [2]. Here we use a simple detector based on kurtosis. Kurtosis is often used as a contrast function reflecting (non)-Gaussian character of a random variable, because the kurtosis of a Gaussian variable is equal to zero. For example, a VAD using kurtosis was proposed in [44]; a recent method for blind source extraction using kurtosis was proposed in [45].

For a complex-valued random variable X , normalized kurtosis is defined as

$$\text{kurt}(X) = \frac{\mathbb{E}[|X|^4] - |\mathbb{E}[X^2]|^2}{\mathbb{E}[|X|^2]^2} - 2, \quad (38)$$

where $\mathbb{E}[\cdot]$ stands for the expectation operator, which is replaced by the sample mean in practice. Speech signals often yield positive kurtosis. We therefore define the set of selected frequencies as

$$k \in \mathcal{S}_{\beta}^{\text{kurt}} \iff \text{kurt}(X_{\text{L}}(\theta_k, \ell)) > \beta. \quad (39)$$

In other words, frequencies that yield higher kurtosis than α on the left channel are supposed to contain a dominating target (speech) signal.

C. Selection Methods after applying BSS

1) *Divergence*: Some BSS methods, such as GSS described in Section III-B.2, proceed by numerical optimization of a contrast function that evaluates the independence of separated outputs. For example, GSS minimizes a criterion for approximate joint diagonalization of covariance matrices of the input signals computed on frames, plus a penalty function ensuring a constraint [30]. When the minimum of the function is shallow, the convergence is slow, which might be indicative of poor separation.

Therefore, the method proposed here rejects frequencies for which the algorithm did not converge within a selected number of iterations. Thus, the selection is

$$k \in \mathcal{S}_Q^{\text{div}} \iff \mathbf{W}(\theta_k) \text{ converged within } Q \text{ iterations.} \quad (40)$$

2) *Coherence-Based Selection*: Another way to assess the separation quality without knowing the achieved SNR is to compute the coherence function among the separated signals. As the separated signals should be independent, the coherence, defined as

$$\text{coh}(\theta_k) = \frac{|\sum_{\ell} y_1(\theta_k, \ell) \overline{y_2(\theta_k, \ell)}|}{\sqrt{\sum_{\ell} |y_1(\theta_k, \ell)|^2} \sqrt{\sum_{\ell} |y_2(\theta_k, \ell)|^2}} \quad (41)$$

should be “small”. Here, $y_i(\theta_k, \ell)$ denotes the i th separated signal, that is, the i th element of $\mathbf{y}(\theta_k, \ell)$ defined in (16). Now, the selection is defined as

$$k \in \mathcal{S}_{\beta}^{\text{coh}} \iff \text{coh}(\theta_k) < \beta. \quad (42)$$

D. Thresholds

Note that there is no clear correspondence between the values of β s in (37), (39) and (42). Rather than determining values for these parameters, β s will be chosen based on a pre-selected ratio of accepted frequencies in percents (this quantity will later be referred to as *percentage*).

VI. EXPERIMENTS

We present results of experiments evaluating and comparing the ability of several methods to attenuate a target speaker in noisy stereo recordings. Each scenario is simulated using a database⁵ of room impulse responses (RIR) measured in the speech & acoustic lab of the Faculty of Engineering at Bar-Ilan University [46]. The lab is a $6 \times 6 \times 2.4$ m room with variable reverberation time (T_{60} is set, respectively, to 160 ms, 360 ms and 640 ms). The database consists of impulse responses relating eight microphones and a loudspeaker. The microphones are arranged to form a linear array (we use pairs of microphones from the arrangement 3 – 3 – 3 – 8 – 3 – 3 – 3 cm) and the loudspeaker is placed at various angles from -90 to 90° at distances of 1 and 2 m; see the setup depicted in Fig. 2. All computations were done in MatlabTM on a standard PC with four-core processor 2.6 GHz and 8 MB of RAM.

Noise signals are either diffused and isotropic (shortly referred to as omnidirectional) or simulated to be directional (one channel of an original noise signal is convolved with RIRs corresponding to the interferer’s position). Sample of omnidirectional babble noise is taken from the database recorded in the lab. Signals for directional sources are taken from the task of the SiSEC 2013 evaluation campaign [47]⁶ titled “Two-channel mixtures of speech and real-world background noise.” We use a female and a male utterance and a sample of babble noise recorded in a cafeteria⁷. The signals are 10 s long, and the sampling frequency is 16 kHz.

⁵<http://www.eng.biu.ac.il/gannot/downloads/>

⁶<http://sisec.wiki.irisa.fr/>

⁷This sample is used to simulate a directional babble noise although typical babble noise is diffused and isotropic. The purpose of this sample is to also have another directional source besides the Gaussian noise.

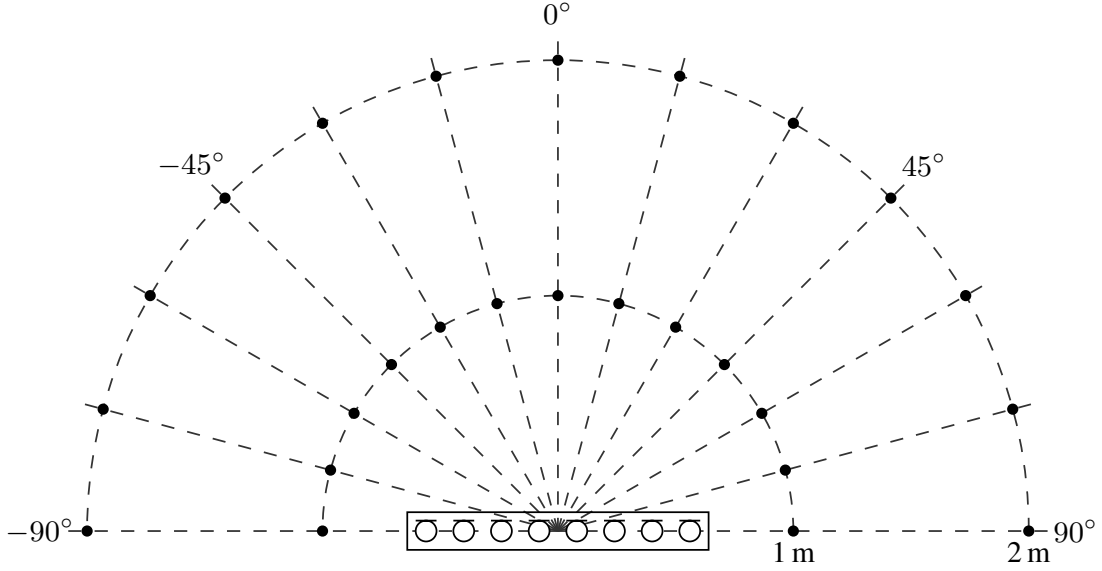


Fig. 2. Illustration of the geometric setup of impulse response database from [46]. The picture is a reprint from [46] with the permission of its authors.

Once microphone responses of the sources are prepared, they are mixed together at a specified SNR averaged over both microphones. Specifically,

$$\text{SNR}_{\text{in}} = \frac{\sum_{i \in \{L, R\}} \sum_n [\{h_i * s\}(n)]^2}{\sum_{i \in \{L, R\}} \sum_n [y_i(n)]^2}, \quad (43)$$

where n spans a given interval of data.

The testing sample (10 s) is split into intervals with 75% overlap; experiments are always conducted on each interval (37 independent trials when the interval length is 1 s) and the results are averaged. For a particular interval, SNR at the output of (6) is computed as

$$\text{SNR}_{\text{out}} = \frac{\sum_n [\{g * s_L\}(n) - s_R(n)]^2}{\sum_n [\{g * y_L\}(n) - y_R(n)]^2}, \quad (44)$$

where $s_R = h_R * s$ (the response of the target signal on the right microphone), and g denotes the estimate of h_{rel} . The numerator of (44) corresponds to the leakage of the target signal in (6) while the denominator contains the desired noise reference.

The final criterion is the *attenuation rate* evaluated as the ratio between SNR_{out} and SNR_{in} . The more negative the value (in dBs) of this criterion is, the better the evaluated filter performs.

We compare several variants of the proposed method combining different approaches to solve Part 1 and Part 2; Part 3 is the same in all instances. The methods used in Part 1 (FD, NSFD and GSS) are always compared with those obtained after Parts 2 and 3, as the main goal is that the latter improve the former; see the list of compared methods in Table I.

If not specified otherwise, parameters are set to the default values shown in Table II. Note that microphone distances are differently selected for FD and NSFD and for GSS in order to provide setups that are preferable for each method (optimized based on the results).

A. Attenuation Rate vs. Percentage

The number of selected frequencies within Part 2 (the parameter we refer to as the *percentage*) has a particular influence on the resulting estimator⁸. On the one hand, the attenuation rate is always poor when

⁸Results of methods that do not allow the choice of the percentage are in graphs shown as constant lines.

TABLE I
METHODS COMPARED IN EXPERIMENTS

Method Acronym	Method used in Part 1	Method used in Part 2
FD	freqv.-dom. estimator (12)	-
FD ^{or}	freqv.-dom. estimator (12)	oracle selector (37)
FD ^{kurt}	freqv.-dom. estimator (12)	kurtosis-based selector (39)
NSFD	non-stationarity-based freqv.-dom. estimator (15)	-
NSFD ^{or}	non-stationarity-based freqv.-dom. estimator (15)	oracle selector (37)
NSFD ^{kurt}	non-stationarity-based freqv.-dom. estimator (15)	kurtosis-based selector (39)
GSS	BSS estimator (17)	-
GSS ^{div}	BSS estimator (17)	divergence-based selector (40)
GSS ^{coh}	BSS estimator (17)	coherence-based selector (42)

TABLE II
DEFAULT SETTINGS IN EXPERIMENTS

Parameter	Value [units]
Sampling frequency	16 kHz
Data interval length per trial	1 s
T ₆₀	360 ms
SNR _{in}	0 dB
Target angle	0°
Directional interferer angle	-60°
Distance of sources to microphones	2 m
Length of DFT M	2048
Window shift in short-term DFT	64
Delay parameter D	100
Microphone pair when using FD or NSFD	[3 4] (3 cm)
Microphone pair when using GSS	[4 5] (8 cm)
c_1, c_2, c_3 in (27)	0.1, 0.11, 0.3
Frame length in NSFD	1000
Number of blocks in GSS	4
$\alpha_{\min}, \alpha_{\max}, \text{tol}$	$10^{-7}, 10^3, 10^{-3}$

the percentage is lower than a certain threshold (depending on the method and experiment). On the other hand, the rate is always getting back to that of the initial estimator as the percentage approaches 100%. It is desirable that the rate should be improved, at least for some values in between these two extremes.

1) *Diffused and isotropic noise*: Figures 3(a) and 3(b) show results from two experiments when the target signal (female speech) is contaminated, respectively, by stationary Gaussian white noise that is spatially white (independently generated on each channel) and by the omnidirectional babble noise.

The white noise situation (Fig. 3(a)) favors NSFD as it obeys the assumed model [6]. Now NSFD^{or} and NSFD^{kurt} perform approximately the same as NSFD or marginally improve the attenuation rate (maximum by 1 dB) unless the percentage goes below 15%. The methods based on FD behave similarly but do not outperform those based on NSFD. The original NSFD is hard to outperform in this scenario as its performance is close to optimal.

In babble noise, NSFD attenuates the target by about 5 dB, while FD yields an attenuation rate above 0 dB, and hence fails. The proposed methods successfully improve these results for a wide range of the percentage values. The best improvements are achieved through oracle methods NSFD^{or} (70%) and FD^{or} (20–80%), where the attenuation rates by NSFD and FD are improved by about 6 dB. The optimum improvement by the kurtosis-based variants NSFD^{kurt} (45%) and FD^{kurt} (45%) is by 4–6 dB, which is only reasonably lower compared to that of the oracle-based frequency selections. The results confirm that the kurtosis-based selection is efficient in detecting frequencies with high SNR when the noise is Gaussian or babble. Examples of estimated ReIRs in this experiment are shown in Fig. 4.

We also examined the case when the target source was shifted to a 60° angle. The results, not shown here due to space constraints, were comparable with the results for 0°.

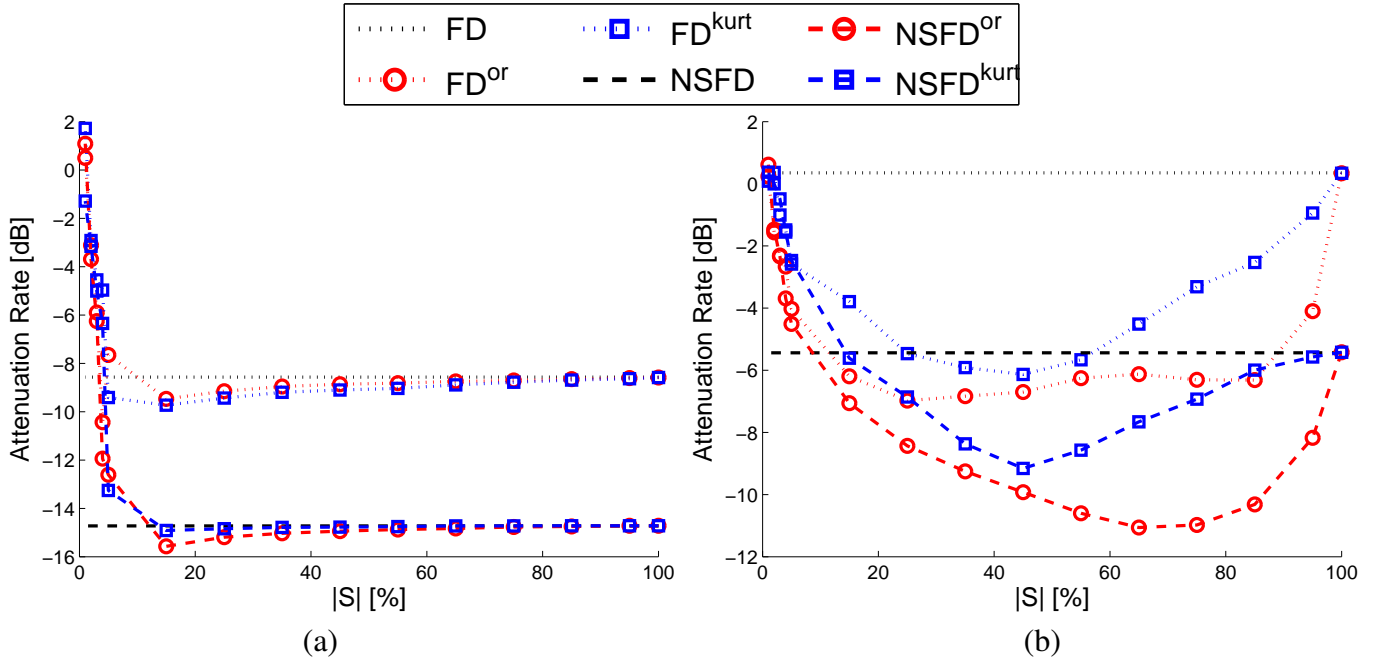


Fig. 3. Female target speaker interfered by (a) Gaussian stationary and spatially and temporally white noise and (b) omnidirectional babble noise.

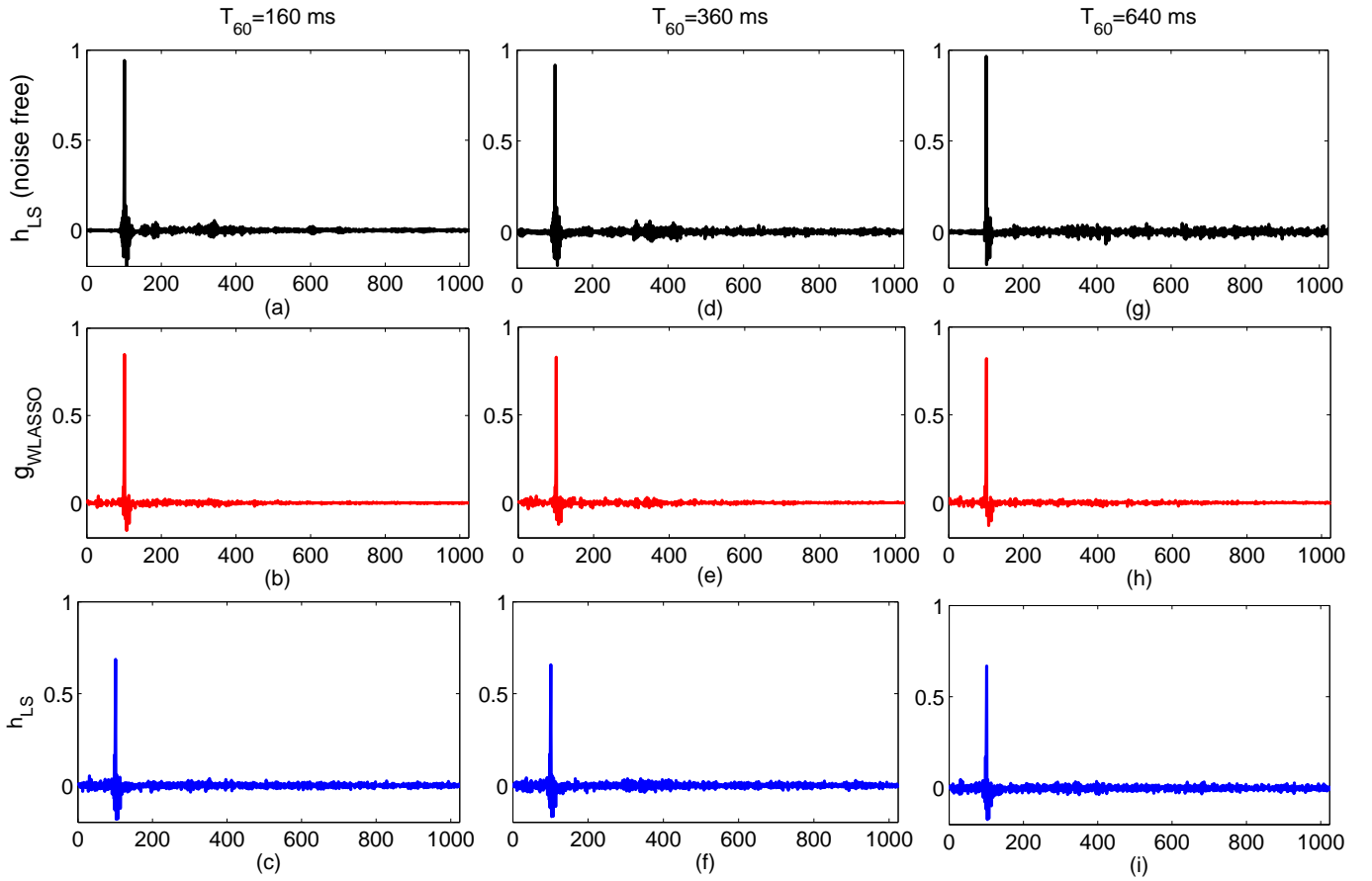


Fig. 4. Examples of ReIRs computed in the first trial of the experiment of Section VI-A for three different reverberation times (columns) when the female target speaker was interfered by omnidirectional babble noise. The first row contains the least-squares estimates according to (9) from noise-free recording of the target while the third row contains the estimates computed from noisy data. The second row contains the sparse approximations computed from 50% incomplete RTF estimate by $NSFD^{kurt}$ (from noisy data). The attenuation rates by the estimated ReIRs were, respectively, (a) -22.4 dB, (b) -12.8 dB, (c) -8.6, (d) -23.7 dB, (e) -11.0 dB, (f) -8.0 dB, (g) -14.7 dB, (h) -7.1 dB, and (i) -6.5 dB.

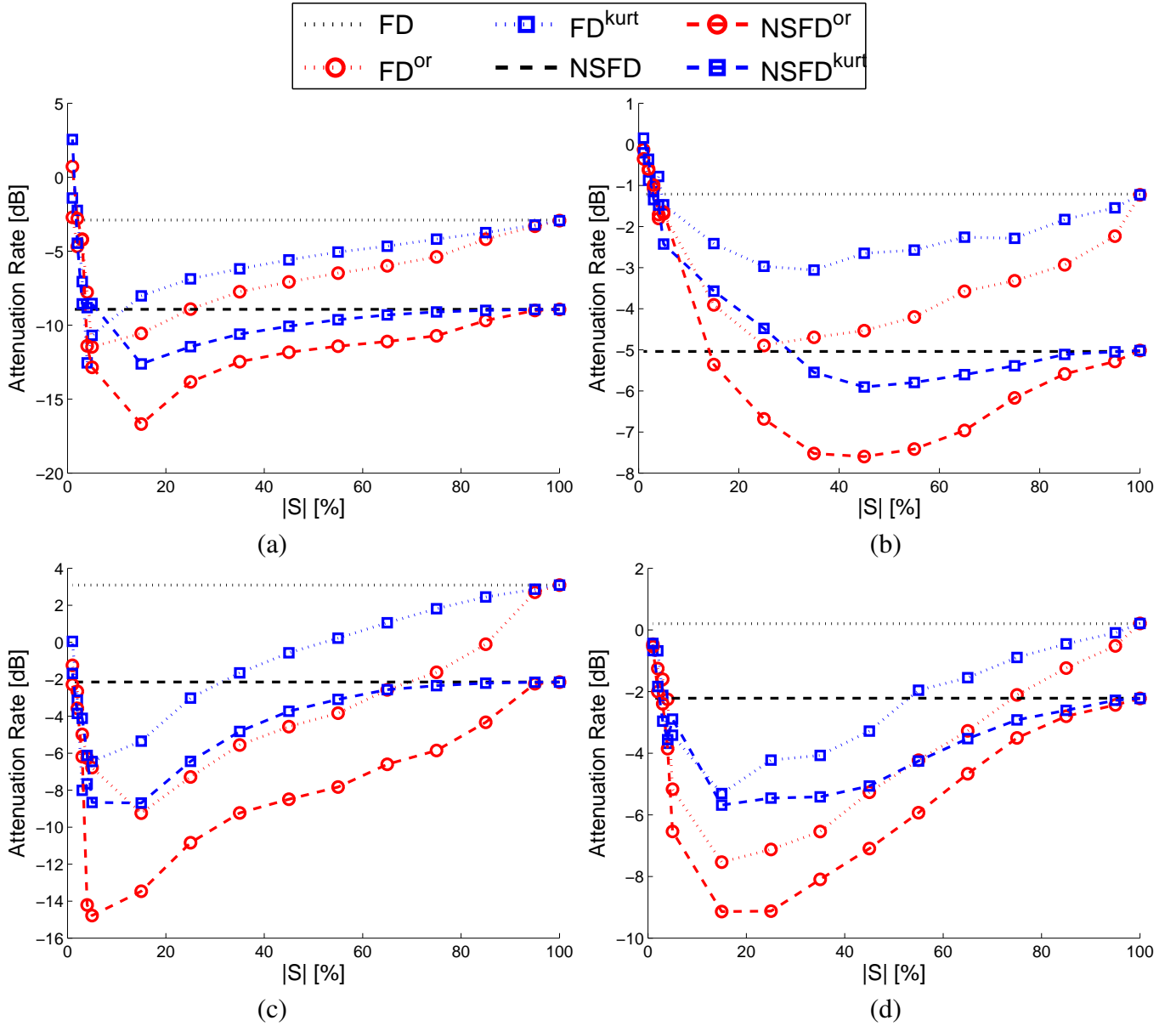


Fig. 5. Results of the experiment where the target source at angle β is interfered by directional noise from -60° : (a) Gaussian noise and $\beta = 0^\circ$, (b) babble noise and $\beta = 0^\circ$, (c) Gaussian noise and $\beta = 60^\circ$, (d) babble noise and $\beta = 60^\circ$.

2) *Directional noise*: Fig. 5 shows results of experiments when noise signals were played from a loudspeaker placed at -60° and the target was placed at an angle of 0° or 60° .

By comparing Fig. 3(a) with Figures 5(a) and 5(c), FD and NSFD perform worse by 5–6 dB and by 11–13 dB, respectively, when the Gaussian noise is directional and the target speaker stands at angles of 0° and 60° . This means the directional noise scenario is now less favorable for both FD and NSFD than in the previous scenario. To explain, note that within the frequency bins with low activity of the target source, these methods, in fact, estimate the RTF of the (directional) noise source. When applying such estimated RTF to attenuate the target signal, part of the noise source is attenuated as well, which causes loss in terms of the attenuation rate.

It should also be noted that the performance loss may be even higher when the target is spatially more separated from the noise source (60°), because the higher the spatial separation of the directional noise source, the higher the bias in the RTF estimates by FD and NSFD could be.

NSFD^{or} and NSFD^{kurt} as well as FD^{or} and FD^{kurt} improve their initial methods, especially when the

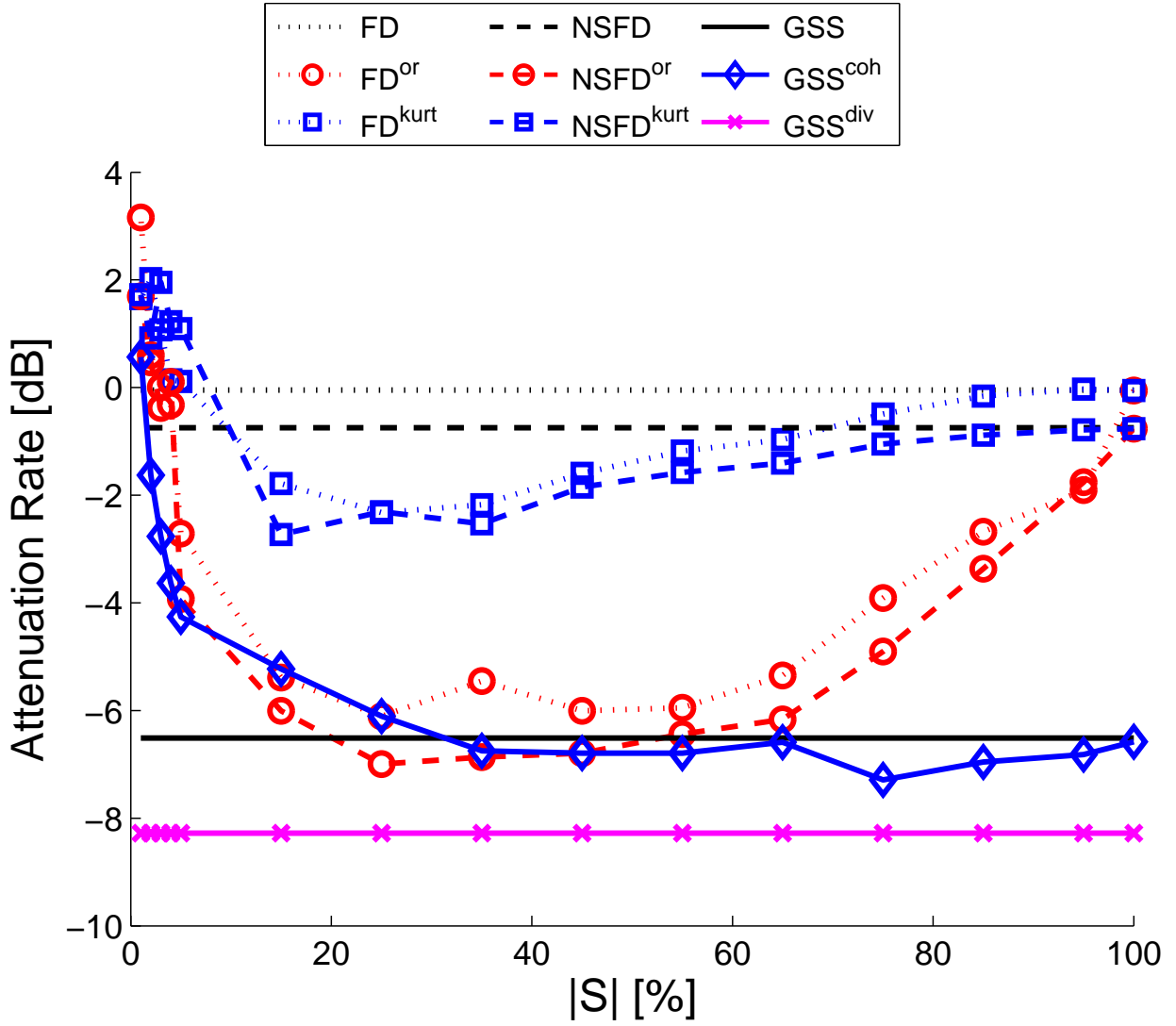


Fig. 6. Female target speaker at 60° interfered with by a male speaker from the angle of -60° , both at the distance of 1 m.

percentage value approaches 15%. Moreover, these methods yield an attenuation rate that is close to that achieved with the spatially white Gaussian noise in Fig. 3(a). Compared to FD and NSFD, the proposed methods do not attenuate the directional noise in the frequency bins with low target source activity.

Similar, but not identical, conclusions can be drawn for the babble noise case. The results by NSFD in Fig. 5(b) are almost the same as those in Fig. 3(b), while, in Fig. 5(d), the attenuation by NSFD drops by 3 dB compared to Fig. 3(b).

3) *A speaking interferer*: A more difficult situation occurs when the interference is a speech signal. We demonstrate this in an experiment where a male speech (interferer) impinges the microphones from the direction of -60° , while a female speaker (target loudspeaker) is placed at 60° ; both at a distance of 1 m; T_{60} is 160 ms. The results are shown in Fig. 6.

Compared to previous experiments, the interfering signal here has similar dynamics and kurtosis as the target signal, which violates the prerequisites of NSFD and of the kurtosis-based selection procedure. Neither FD, NSFD nor FD^{kurt} and $NSFD^{kurt}$ can distinguish the target speaker from the interfering one and, therefore, all of them perform much worse than FD^{or} and $NSFD^{or}$ (for a large range of percentage values).

By looking closer at FD and NSFD, they actually try to attenuate both signals by eliminating the dominating signal within each frequency. To show this fact, we performed a simple experiment by taking

only the first trial interval of this experiment. Here, FD^{or} and NSFD^{or} achieved, respectively, attenuation rates of -7.0 and -7.42 dB with a percentage of 25%. When the roles of the target and interfering speaker were interchanged so that the oracle procedures took 25% of frequencies where the interferer was dominant, FD^{or} and NSFD^{or} attenuated the interferer, respectively, by 11.3 and 11.2 dB. The fact that both results were obtained from the same RTF estimates by just selecting different frequency bins confirms that FD and NSFD tend to attenuate both signals.

In this experiment, we further consider GSS which is capable of blindly separating the target signal from the interference and vice versa⁹. The RTF estimate can be obtained as described in Section III-B.2. Then we can also apply the proposed method based on the selection procedures (40) and (42).

The results in Fig. 6 show that GSS outperforms NSFD as well as FD. Next, GSS^{div} (here with $Q = 30$) attenuates the target by about 8 dB, which improves GSS by 2 dB. Here GSS^{coh} also improves the attenuation rate achieved by GSS, where the best improvement is achieved for 70–80%. Hence, GSS^{div} appears to be better than GSS^{coh} . However, other experiments not shown here due to space limitations prove that this comparison does not hold in general.

B. Attenuation Rate versus Length of Data

Fig. 7 shows results of repeated experiments, respectively, with temporally and spatially white Gaussian noise and omnidirectional babble noise. The selection percentage of the proposed methods was, respectively, fixed at 25% and 45% while the data length was varied from 250 ms to 2 s.

The attenuation rates of FD and NSFD are slowly improved with a growing interval length. Also the performance of the proposed variants is improved with a growing length of data. On the other hand, the improvement is not necessarily monotonic, since the attenuation rate also depends on the percentage, which is fixed in this experiment. An example of the non-monotonic performance is that of NSFD^{or} in Fig. 7(b). Next, $\text{NSFD}^{\text{kurt}}$ and FD^{kurt} perform even worse than NSFD and FD for the data length of 250 ms. This may be solved by increasing the percentage in the latter methods closer to 100%. The performances of NSFD^{or} and FD^{or} remain stable for all data lengths, which points to room for possible improvements (e.g. more robust selection procedures).

C. Varying SNR_{in}

Here, the experiments where the babble noise was played from a loudspeaker (Fig. 5(b)) and with the male interferer (Fig. 6) are, respectively, repeated with the percentage fixed, respectively, at 45% and 55%; SNR_{in} was changed from -10 to 10 dB. Fig. 8 shows the resulting attenuation rates.

The performance of FD and NSFD is improving with growing SNR_{in} . For SNR_{in} below about 0 dB, their attenuation rate goes above zero, because the interfering source is becoming dominant, and FD and NSFD aim to attenuate the former more than the target signal.

The proposed methods achieve a better attenuation rate than FD and NSFD for almost all values of SNR_{in} . An exception occurs when $\text{SNR}_{\text{in}} = 10$ dB. Here, $\text{NSFD}^{\text{kurt}}$ (and also NSFD^{or} in Fig. 8(a)) perform worse than NSFD. This is again due to the fixed percentage value, which should be chosen close to 100% when SNR_{in} is high. For $\text{SNR}_{\text{in}} = 10$ dB, NSFD appears to be efficient.

In the experiment of Fig. 8(b), GSS and the variants derived therefrom perform almost constantly and are only slightly improved with the growing SNR_{in} . This is due to the blind separation of the sources by GSS, which is very efficient when sources are closer to microphones (1 m here) and the reverberation time is low ($T_{60} = 160$ ms).

D. Varying T_{60}

The last experiment considers varying reverberation time when T_{60} is respectively 160, 360 and 640 ms (the values available in the database [46]). The experiment with two speakers is repeated here with the percentage fixed at 55%.

⁹We apply GSS using known DOAs in this experiment.

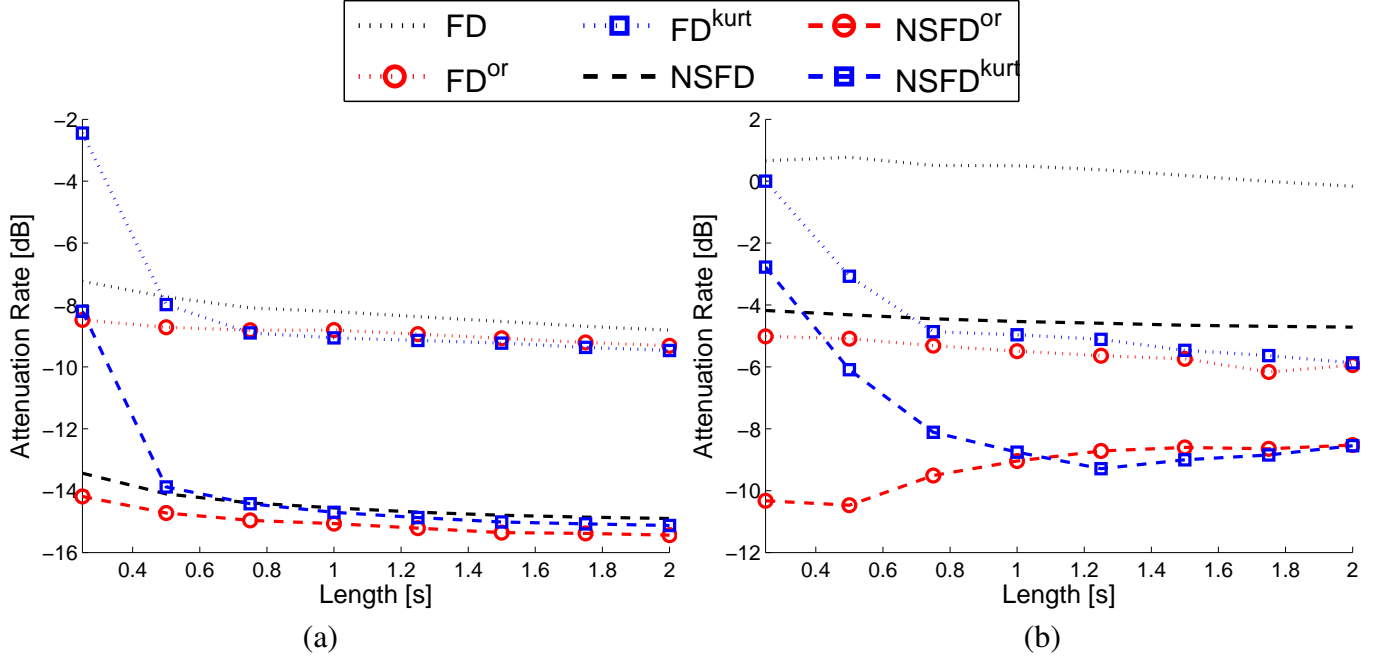


Fig. 7. Dependence of attenuation rate on the length of data interval. The target speaker is interfered with by (a) temporally and spatially white Gaussian noise and (b) omnidirectional babble noise.

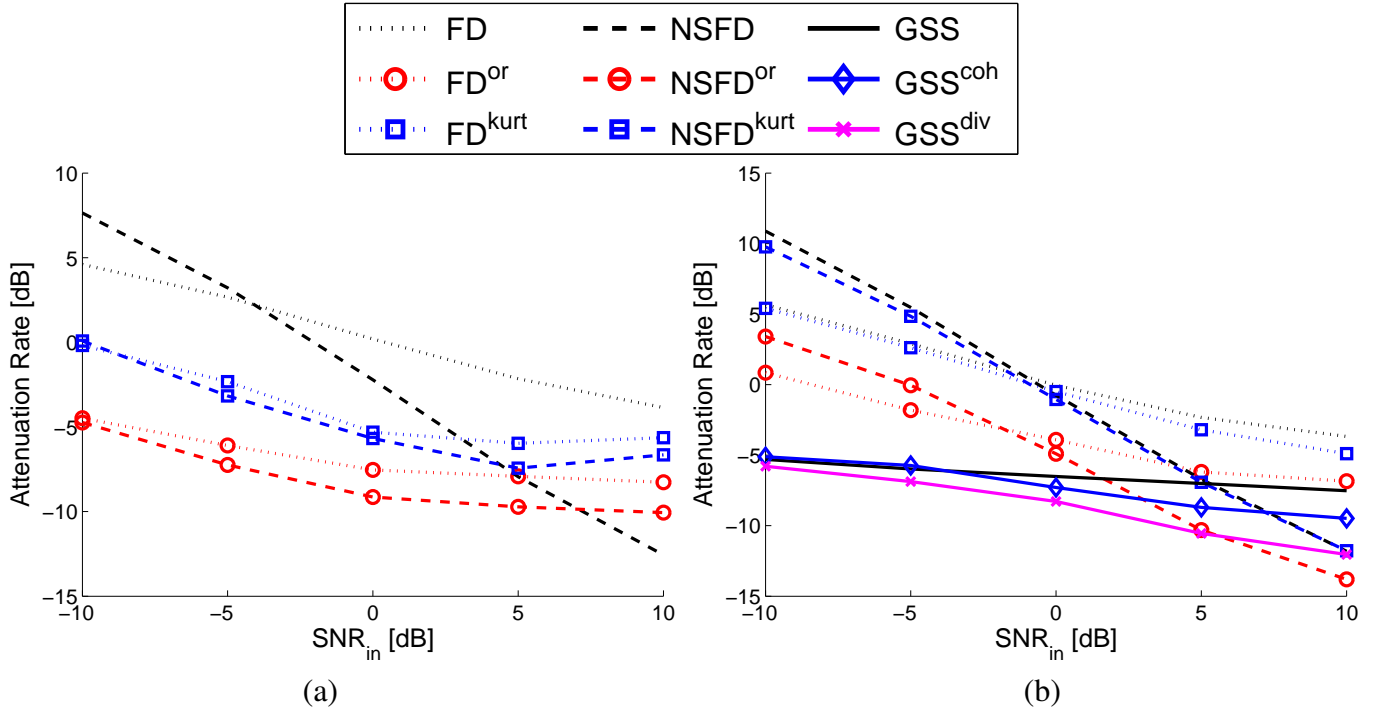


Fig. 8. Attenuation rate as a function of SNR_{in} when the target's angle is 60° and the noise is (a) directional babble coming from a 0° angle and (b) male speech coming from a -60° angle.

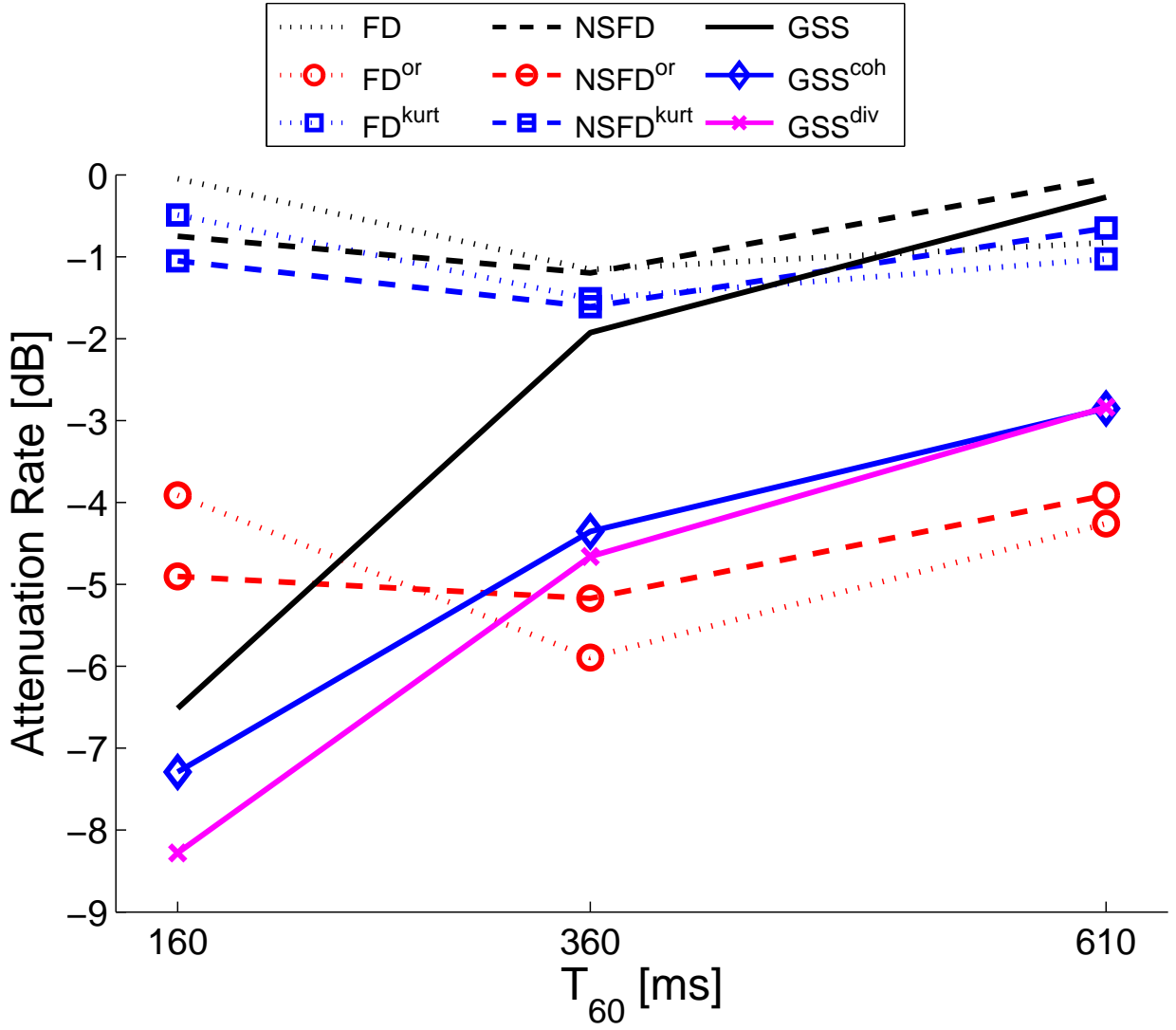


Fig. 9. Attenuation rates as functions of reverberation time T_{60} . Female target voice at 60° was interfered with a male voice played from the angle of -60° both at the distance of 1 m; $\text{SNR}_{\text{in}} = 0$ dB.

FS, NSFD and their kurtosis-based variants do not succeed here for any value of T_{60} for the same reasons as in the experiment of Section VI-A.3. By contrast, the attenuation rates of NSFD^{or} and FD^{or} are only slightly dependent on T_{60} , which points to the necessity to distinguish the target's and interferer's frequencies correctly. The performance of FD^{or} is even improving with T_{60} , but this is again due to the fixed percentage whose optimum value is different for each situation.

The attenuation rate by GSS, GSS^{div} and GSS^{coh} is dropping as the T_{60} is growing, because the blind separation is becoming difficult with the reverberation time of the environment. Nevertheless, both GSS^{div} and GSS^{coh} improve the attenuation rate by GSS up to by 3 dB even in the most difficult case when $T_{60} = 640$ ms.

VII. CONCLUSIONS AND DISCUSSION

We have proposed a novel approach estimating the RTF from noisy data. The experiments have shown that, in most situations, the proposed approach yields RTF estimates better than conventional estimators in terms of the capability to cancel the target signal. The crucial parameter to select is the percentage. The optimum percentage depends on many circumstances and is hard to predict. Nevertheless, the experiments where the percentage was fixed have shown that the performance of the method is not too sensitive to

this parameter. The performance gain due to the method remains positive when reasonable percentage is chosen, e.g., based on practice.

The proposed method is flexible in providing room for future modifications and improvements, some of which we list now.

Methods for solving particular parts of the method can be replaced by novel ones, especially the conventional estimators used within the first part. The methods could be tailored to particular scenarios, signal mixtures or noise conditions. For example, we have demonstrated through experiments that NSFD is effective for the first part when noise is isotropic and less dynamic than the target speech signal, while GSS can be efficient when noise is a competitive speech signal.

If some prior knowledge of SNR (or other knowledge) is available, the selection of frequencies (the second part) could be done before or simultaneously with the RTF estimation (the first part). This could lead to computational savings as only the incomplete RTF estimate needs to be computed.

In the method proposed here, the RTF estimate is reconstructed through searching for the sparsest representation of the incomplete RTF in the discrete time-domain. Besides the fact that faster methods for solving (26) may appear in the future, the weighted ℓ_1 program is by far not the only way to reconstruct the RTF estimate [48]. For example, it is possible to reconstruct the RTF in an over-sampled discrete time-domain or in the continuous time-domain; see [49], [50].

Online or batch-online implementations of the proposed methods can be the subject of future developments. For each part, it is possible to select an appropriate online or adaptive method to solve the corresponding task.

REFERENCES

- [1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.
- [2] I. Tashev, *Sound Capture and Processing: Practical Approaches*, John Wiley & Sons Ltd., 2009.
- [3] S. Gannot and I. Cohen, "Springer Handbook of Speech Processing and Speech Communication," ch. "Adaptive beamforming and postfiltering," New York: Springer-Verlag, 2007.
- [4] J. Benesty, S. Makino, and J. Chen (Eds.), *Speech Enhancement*, 1st edition, Springer-Verlag, Heidelberg, 2005.
- [5] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.
- [6] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.
- [7] S. Affes, Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 425–437, Sept. 1997.
- [8] A. Krueger, E. Wartsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 206–219, Jan. 2011.
- [9] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.
- [10] S. Markovich, S. Gannot and I. Cohen, "Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment with Multiple Interfering Speech Signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.
- [11] K. Yen and Y. Zhao, "Adaptive Co-Channel Speech Separation and Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 138–151, March 1999.
- [12] J.-F. Cardoso, "Blind signal separation: statistical principles", *Proceedings of the IEEE*, vol. 90, n. 8, pp. 2009–2026, October 1998.
- [13] F. Nesta and M. Omologo, "Convolutional underdetermined source separation through weighted interleaved ICA and spatio-temporal source correlation," *Proc. of The 10th International Conference on Latent Variable Analysis and Source Separation (LVA/ICA 2012)*, pp. 222–230, Tel-Aviv, Israel, March 12–15, 2012.
- [14] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, W. Kellermann, "Geometrically Constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–4, 2013.
- [15] Z. Koldovský, P. Tichavský, D. Botka, "Noise Reduction in Dual-Microphone Mobile Phones Using A Bank of Pre-Measured Target-Cancellation Filters," *Proc. of ICASSP 2013*, pp. 679–683, Vancouver, Canada, May 2013.
- [16] Z. Koldovský, J. Málek, P. Tichavský, and F. Nesta, "Semi-blind Noise Extraction Using Partially Known Position of the Target Source," *IEEE Trans. on Speech, Audio and Language Processing*, vol. 21, no. 10, pp. 2029–2041, Oct. 2013.
- [17] R. Talmon and S. Gannot, "Relative transfer function identification on manifolds for supervised GSC beamformers," in *Proc. of the 21st European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, Sep. 2013.
- [18] Y. Lin, J. Chen, Y. Kim and D. Lee, "Blind channel identification for speech dereverberation using ℓ_1 norm sparse learning," *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems*, MIT Press, Vancouver, British Columbia, Canada, December 3–6, 2007.

- [19] M. Yu, W. Ma, J. Xin, S. Osher, "Multi-Channel l_1 Regularized Convex Speech Enhancement Model and Fast Computation by the Split Bregman Method," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 661–675, Feb. 2012.
- [20] J. Málek and Z. Koldovský, "Sparse Target Cancellation Filters with Application to Semi-Blind Noise Extraction," *Proc. of the 41st IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, pp. 2109–2113, May 2014.
- [21] A. Benichoux, L. S. R. Simon, E. Vincent and R. Gribonval, "Convex Regularizations for the Simultaneous Recording of Room Impulse Responses," *IEEE Transactions on Signal Processing*, vol. 62, no. 8, pp. 1976–1986, April 2014.
- [22] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.
- [23] O. Hoshuyama, A. Sugiyama, A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.
- [24] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, K. Shikano, "Blind Spatial Subtraction Array for Speech Enhancement in Noisy Environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 650–664, May 2009.
- [25] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, W. Kellermann "A Stereophonic Acoustic Signal Extraction Scheme for Noisy and Reverberant Environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 726–745, May 2012.
- [26] E. Habets and S. Gannot, "Dual-microphone speech dereverberation using a reference signal," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, vol. 4, no. IV, pp. 901–904, Apr. 2007.
- [27] N. Levinson, "The Wiener RMS error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, pp. 261–278, 1947.
- [28] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second-order statistics: A time domain approach," *IEEE Trans. Information Theory*, vol. 40, no. 2, pp. 340–349, 1994.
- [29] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 2055–2063, Aug. 1996.
- [30] L. C. Parra and Ch. V. Alvino, "Geometric Source Separation: Merging Convolutional Source Separation With Geometric Beamforming," *IEEE Trans. on Signal Processing*, vol. 10, no. 6, Sept. 2002.
- [31] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no.2, pp. 666–678, March 2006.
- [32] E. J. Candès and T. Tao, "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.
- [33] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
- [34] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Communications on Pure and Applied Mathematics*, vol. 61, no. 8, pp. 1025–1045, August 2008.
- [35] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
- [36] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, May 2009.
- [37] S. S. Chen, D. L. Donoho, M. A. Saunders, "Atomic Decomposition by Basis Pursuit", *SIAM Journal on Scientific Computing*, Vol. 20, No. 1., pp. 33–61, 1999.
- [38] E. van den Berg and M. P. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM J. on Scientific Computing*, vol. 31, no. 2, pp.890–912, Nov. 2008.
- [39] D. L. Donoho, Y. Tsaig, "Fast Solution of l_1 -Norm Minimization Problems When the Solution May Be Sparse", *IEEE Transactions on Information Theory*, Vol. 54, No. 11., pp. 4789–4812, 2008.
- [40] S. J. Wright, R. D. Nowak, M. A. T. Figueiredo, "Sparse Reconstruction by Separable Approximation," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, July 2009.
- [41] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *SIAM J. Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.
- [42] N. Parikh and S. Boyd, "Proximal Algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, Nov. 2013.
- [43] M. S. Asif and J. Romberg, "Fast and accurate algorithms for re-weighted L_1 -norm minimization," *submitted to IEEE Trans. on Signal Process.*, arXiv:1208.0651, July 2012.
- [44] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217–231, March 2001.
- [45] S. Javidi, D. P. Mandic, C. C. Took, and A. Cichocki, "Kurtosis-based blind source extraction of complex non-circular signals with application in EEG artifact removal in real-time," *Frontiers in Neuroscience*, vol. 5, no. 105, 2011.
- [46] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," *International Workshop on Acoustic Signal Enhancement 2014 (IWAENC 2014)*, Antibes, France, Sept. 2014.
- [47] N. Ono, Z. Koldovský, S. Miyabe, N. Ito, "The 2013 Signal Separation Evaluation Campaign," *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, Southampton, UK, Sept. 2013.
- [48] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky, "The Convex Geometry of Linear Inverse Problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.
- [49] B. N. Bhaskar, T. Gongguo, and B. Recht, "Atomic Norm Denoising With Applications to Line Spectral Estimation," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5987–5999, Dec. 2013.
- [50] Z. Koldovský and P. Tichavský, "Sparse Reconstruction of Incomplete Relative Transfer Function: Discrete and Continuous Time Domain," *submitted to a special session of EUSIPCO 2015*, Feb. 2015.



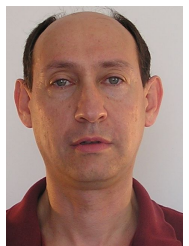
Zbyněk Koldovský (S'03-M'04) was born in Jablonec nad Nisou, Czech Republic, in 1979. He received the M.S. degree and Ph.D. degree in mathematical modeling from Faculty of Nuclear Sciences and Physical Engineering at the Czech Technical University in Prague in 2002 and 2006, respectively.

He is currently an associate professor at the Institute of Information Technology and Electronics, Technical University of Liberec. He has also been with the Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic since 2002. His main research interests are focused on audio signal processing, blind source separation, statistical signal processing, compressed sensing, and multilinear algebra.

Dr. Koldovský serves as a reviewer for several journals such as the IEEE Transaction on Audio, Speech, and Language Processing, IEEE Transaction on Signal Processing, Elsevier Signal Processing Journal, and in several conferences and workshops in the field of (acoustic) signal processing. He has served as a co-chair in the fourth community-based Signal Separation Evaluation Campaign (SiSEC 2013) and as the general co-chair of the twelfth International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2015).



Jiří Málek received his master and Ph.D. degrees from Technical University in Liberec (TUL, Czech Republic) in 2006 and 2011, respectively, in technical cybernetics. Currently, he holds a postdoctoral position at the Institute of Information Technology and Electronics, TUL. His research interests include blind source separation and speech enhancement.



Sharon Gannot (S'92-M'01-SM'06) received his B.Sc. degree (summa cum laude) from the Technion Israel Institute of Technology, Haifa, Israel in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Israel in 1995 and 2000 respectively, all in electrical engineering. In 2001 he held a post-doctoral position at the department of Electrical Engineering (ESAT-SISTA) at K.U.Leuven, Belgium. From 2002 to 2003 he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. Currently, he is an Associate Professor at the Faculty of Engineering, Bar-Ilan University, Israel, where he is heading the Speech and Signal Processing laboratory. Prof. Gannot is the recipient of Bar-Ilan University outstanding lecturer award for 2010 and 2014.

Prof. Gannot has served as an Associate Editor of the EURASIP Journal of Advances in Signal Processing in 2003-2012, and as an Editor of two special issues on Multi-microphone Speech Processing of the same journal. He has also served as a guest editor of ELSEVIER Speech Communication and Signal Processing journals. Prof. Gannot has served as an Associate Editor of IEEE Transactions on Speech, Audio and Language Processing in 2009-2013. Currently, he is a Senior Area Chair of the same journal. He also serves as a reviewer of many IEEE journals and conferences. Prof. Gannot is a member of the Audio and Acoustic Signal Processing (AASP) technical committee of the IEEE since Jan., 2010. He is also a member of the Technical and Steering committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) since 2005 and was the general co-chair of IWAENC held at Tel-Aviv, Israel in August 2010. Prof. Gannot has served as the general co-chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in October 2013. Prof. Gannot was selected (with colleagues) to present a tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013 and EUSIPCO 2013. Prof. Gannot research interests include multi-microphone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation; dereverberation; single microphone speech enhancement and speaker localization and tracking.